Jurnal Teknologi Maklumat & Multimedia **5**(2009): 47 -60

Pengelasan Dokumen Web di Bursa Malaysia Menggunakan Pendekatan *Support Vector Machine* (SVM)

MOHD SHAHIZAN OTHMAN, LIZAWATI MI YUSUF,
JUHANA SALIM & ZARINA SHUKUR

ABSTRAK

Perkhidmatan Internet yang semakin mudah dicapai telah membawa kepada peningkatan bilangan laman web yang drastik. Fenomena ini telah menimbulkan masalah baru untuk enjin carian dan penstrukturan sumber maklumat web. Enjin carian semasa telah didapati menghasilkan terlalu banyak maklumat yang tidak relevan dan pautan yang hilang. Direktori Internet dan carian topik yang khusus pula, mampu menghasilkan keputusan yang lebih berkualiti berbanding enjin carian. Walau bagaimanapun, proses pembinaan dan penyelenggaraannya melibatkan kos yang tinggi kerana melibatkan penggunaan tenaga pakar untuk mengelaskan sumber maklumat. Jadi, kaedah pengelasan data yang tepat dan berkesan amat diperlukan untuk membolehkan maklumat yang berkualiti dapat dicapai. Kertas kerja ini membincangkan tentang kajian terhadap pengelasan laman web syarikat di Papan Utama dan Kedua, Bursa Malaysia menggunakan pendekatan Support Vector Machine (SVM). Hasil kajian menunjukkan pengelasan dokumen web menggunakan kernel linear *mencatatkan peratusan ketepatan pengelasan yang terbaik.*

Katakunci: Pengelasan, Support Vector Machine*, Organisasi Maklumat*

ABSTRACT

Internet services which has been more easily accesible has drastically brought an increased amount of web sites. The phenomena relating to the large increase in the size of web sites creates new problems for search engines and the stucturing of web information resources. Present search engines are found to be providing too many lost links and information that are not relevant. In contrast to search engines, Internet directories and specific topic search produce better search results. However, the process of development and maintenance of directories involve high cost as it requires the use of experts to classify web sources. Retrieval of quality information requires methods that could classify sources precisely and effectively. This article discusses research relating to the classification of companies' web site in the Main board and Second board, Bursa Malaysia using Support Vector Machine (SVM) method. The result reveals that web classification using linear kernel gives the best percentage of accuracy.

Keywordi: Classification, Support Vector Machine, Information Organization

PENGENALAN

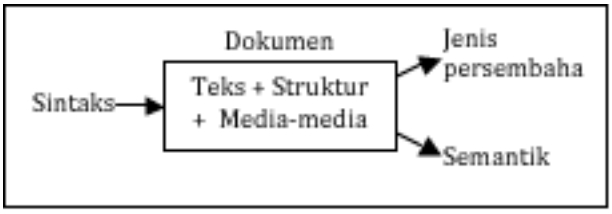
Pengelasan dokumen web merupakan proses yang sistematik untuk menyusun atau mengorganisasikan dokumen web yang semakin bertambah dari masa ke semasa. Beberapa kajian yang telah dijalankan mendapati bahawa terdapat kira-kira 1 bilion laman web yang boleh dicapai dan sebanyak 1.5 juta laman web baru dihasilkan setiap hari. Jumlah laman web ini semakin meningkat dan telah mencapai 11.5 bilion laman web pada bulan Januari 2005 (Lawrence & Giles 1999; Pierre 2001; Gulli & Signorini 2005). Perkembangan sumber maklumat web yang pesat ini, menjadi satu cabaran kepada pengguna Internet untuk mencapai pelbagai maklumat terkini yang relevan dan berkualiti. Menurut Rachagan (2005), melayari laman web dengan kumpulan data yang terlalu banyak menyebabkan ramai individu terpaksa mengambil masa yang lama untuk mencari, mengumpul dan menyusun data. Ini mengakibatkan pengurangan tahap produktiviti mereka. Kajian yang dilaksanakan oleh Rachagan dan disokong oleh Hizral Tazzif (2005), mendapati bahawa secara umumnya, rakyat Malaysia menghabiskan masa selama sembilan jam seminggu bagi melayari Internet dengan pecahan 43% melayari sistem rangkaian komputer itu kurang empat jam seminggu, 24.9% antara empat hingga lapan jam seminggu, 14.2% (antara lapan hingga 15 jam seminggu), 5.7% (antara 15 hingga 22 jam seminggu) dan 3.5% antara 22 hingga 28 jam seminggu. Walaupun banyak masa diperuntukkan untuk mendapatkan maklumat di Internet, tetapi hasilnya sangat mendukacitakan dan mengecewakan pengguna.

Selain itu, fenomena perkembangan web yang pesat ini juga turut mendapat tumpuan komuniti yang membuat kajian tentang pengindeksan web. Menurut Henzinger et al. (2003), isu penting yang dititikberatkan dalam pengindeksan web, antaranya ialah mencari sumber maklumat yang tepat dan relevan, mempercayai kandungan maklumat, melayari sumber maklumat yang baru dan menggunakan kata kunci yang betul untuk mencapai maklumat yang dikehendaki. Hammerich dan Harrison (2002) pula berpendapat, terdapat banyak laman web yang mana isi kandungannya tidak tersusun dan terurus. Keadaan ini menyusahkan pengguna semasa melayari laman web tersebut dan ini sekaligus menjejaskan imej sesebuah syarikat. Jadi, pelbagai usaha perlu diambil oleh pihak tertentu untuk mencari penyelesaian bagi mengelaskan dokumen web.

PENGELASAN DOKUMEN WEB

Penyelidikan dalam pengelasan teks merupakan gabungan bidang perlombongan data, pembelajaran mesin dan capaian maklumat. Kajian tentang pengelasan teks yang pertama telah dilaksanakan oleh Maron (1961), membincangkan tentang pengindeksan automatik. Pada ketika itu, penyelidikan hanya tertumpu kepada pengelasan dokumen teks yang terdapat di perpustakaan sahaja. Namun begitu, kepentingan pengelasan teks telah berkembang dan menjadi penyelidikan utama dalam disiplin sistem maklumat pada awal tahun 90an. Hal ini disebabkan oleh perkembangan teknologi digital terutamanya Internet dan pertambahan bilangan laman web yang begitu pesat menyebabkan kesukaran mengelaskan sumber maklumat web secara manual. Justeru itu, kajian ini dilaksanakan untuk mengenal pasti semua proses yang terlibat dalam pengelasan teks, khususnya dokumen web.

Berbeza dengan dokumen teks biasa, dokumen web mempunyai ciri-ciri tersendiri di mana ianya bukan hanya terdiri daripada teks tetapi merupakan gabungan teks, struktur, audio, video, imej dan tag. Semua gabungan ini digunakan untuk menghasilkan satu laman web. Rajah 1 menunjukkan ciri dokumen web yang sintaks dokumen akan menyatakan struktur, jenis persembahan, semantik ataupun tindakan luaran (Baeza-Yates & Ribeiro-Neto 1999). Sintaks ini terdiri daripada arahan yang lengkap (gabungan tag). Walaupun terdapat 109 tag HTML yang wujud, tetapi bukan semuanya digunakan serentak dalam penghasilan sesebuah dokumen web.



RAJAH 1. Ciri dokumen web (Baeza-Yates & Ribeiro-Neto 1999)

Kajian ke atas struktur dokumen web oleh Etzioni (1996), Kosala dan Blockheel (2000) serta Mohd Shahizan et. al (2005) mendapati kebanyakan dokumen web lebih mementingkan jenis persembahan pada pelayar web berbanding dengan kandungannya. Seterusnya, kajian Mohd Shahizan et. al (2005) mendapati peratusan penggunaan tag adalah lebih tinggi (48.75%) berbanding dengan kandungan dokumen webnya (26.12%). Jadual 1 menunjukkan peratusan penggunaan perkataan bagi dokumen web.

JADUAL 1. Peratusan Penggunaan Perkataan Bagi Dokumen Web

Jenis Kandungan	**Peratus**
Tag	48.75
Kata Henti	21.52
Kandungan	26.12
Imbuhan	3.61

Sumber: Mohd Shahizan et al. 2005

Ini bermakna, pengelasan web merupakan proses yang sistematik untuk menyusun atau menstrukturkan dokumen web. Pengelasan dokumen web juga, merupakan kerja memilih kategori yang berkaitan untuk sesuatu laman web supaya laman web tersebut boleh dimasukkan ke dalam kategori yang sesuai untuknya. Walaupun begitu, menurut Amoretti (2006), dokumen web merupakan masalah dalam proses perlombongan data terutama sekali untuk pengelasan teks. Hal ini disebabkan sesuatu dokumen web terdiri daripada gabungan teks dan tag. Sekiranya pra-pemproses seperti pembuangan tag dan kata henti dilaksanakan, bilangan teks yang tinggal masih lagi banyak. Keadaan ini menjadi tumpuan para penyelidik untuk mencari penyelesaian terbaik bagi mengelaskan dokumen web secara automatik (Sebastiani 2005).

Selain itu, pengelasan dokumen web berbeza dengan pengelasan teks tradisional. Perbezaannya, sifat dokumen web yang tidak berstruktur, mempunyai kadar hingar yang tinggi dan terdiri daripada gabungan pelbagai jenis tag dan teks. Pal et al. (2002) menyatakan yang kriteria dokumen web adalah tidak berlabel, teragih, heterogen, semistruktur, berubah mengikut masa dan berdimensi tinggi. Manakala Huang (2000) pula menyatakan, secara amnya kriteria dokumen web adalah seperti berikut:

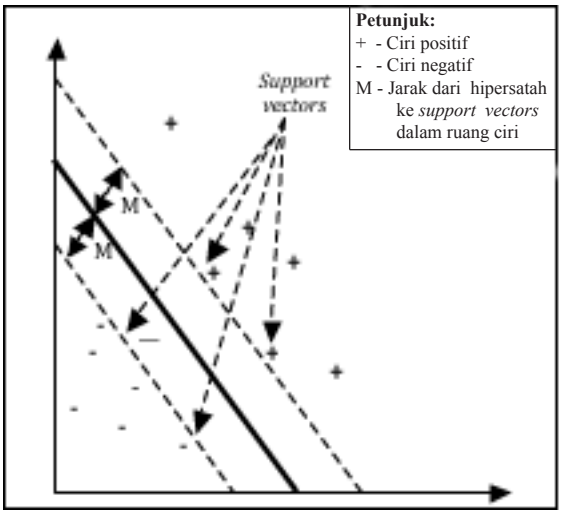
(i) Bersaiz sangat besar: keseluruhan dokumen web berjumlah 350 juta dokumen pada Julai 1998, dan saiz perkembangannya adalah 20 juta dokumen per bulan.
(ii) Bersifat dinamik: dokumen web berubah setiap hari berbanding dengan pangkalan data teks yang statik.
(iii) Heterogen: Internet merangkumi pelbagai jenis dokumen seperti imej, fail audio, teks dan skrip.
(iv) Kepelbagaian bahasa: jumlah keseluruhan penggunaan bahasa di internet melebihi 100 bahasa.
(v) Pautan yang tinggi: setiap dokumen web mempunyai sekurang-kurangnya lapan pautan kepada laman lain.

Jadi, bagi menangani permasalahan dalam pengelasan dan kriteria dokumen web yang pelbagai, terdapat banyak jenis algoritma pengelasan yang telah dibangunkan. Contohnya Bayes Naif, pepohon keputusan, peraturan keputusan, k-*Nearest Neighbour*, *Support Vector Machine* (SVM) dan rangkaian neural. Walau bagaimanapun, artikel ini hanya tertumpu kepada kajian tentang pengelasan dokumen web menggunakan kaedah pengekstrakan dan *Support Vector Machine* (SVM).

KAEDAH *SUPPPORT VECTOR MACHINE* (SVM)

Kaedah SVM telah diperkenalkan dalam bidang pengelasan teks oleh Joachims dan digunakan oleh Dumais et al. (1998), Drucker et al. (1999), Yang dan Liu (1999) serta Klinkenberg dan Joachims (2000). SVM menyediakan dua sifat yang tidak terdapat pada algoritma pembelajaran yang lain iaitu proses memaksimumkan margin dan transformasi ruang input bukan-*linear* kepada ruang ciri menggunakan kaedah kernel (Cortes & Vapnik 1995). Algoritma SVM beroperasi dengan memetakan set latihan yang diberi kepada satu ruang ciri yang berdimensi tinggi dengan menempatkan satu pembahagi yang boleh mengasingkan model positif daripada model negatif dalam ruang tersebut. Dokumen-dokumen yang dilabelkan sebagai positif merupakan dokumen-dokumen yang berada dalam kategori c_i, manakala dokumen-dokumen yang dilabelkan sebagai negatif merupakan dokumen-dokumen yang bukan berada di dalam kategori c_i.

Bentuk SVM yang paling mudah ialah SVM *linear*. SVM *linear* merupakan satu hipersatah iaitu sempadan kelas yang mengasingkan set data positif daripada set data negatif dengan margin yang maksimum di dalam ruang ciri. Margin (*M*) menandakan jarak dari hipersatah ke data positif dan negatif yang terdekat dalam ruang ciri (Yu et al. 2002). Rajah 2 menunjukkan contoh masalah dua-dimensi mudah yang boleh diasingkan secara *linear*. Merujuk kepada Rajah 2, setiap ciri berpadanan dengan satu dimensi di dalam ruang ciri. Jarak dari hipersatah ke satu titik data adalah ditentukan oleh kekuatan setiap ciri dalam data tersebut. Sebagai contoh, satu pengelas dokumen multimedia dipertimbangkan. Sekiranya satu dokumen meliputi banyak ciri yang berkait dengan konsep "multimedia", contohnya perkataan-perkataan "multimedia", "grafik" atau "audio" pada bahagian kepala, dokumen ini akan digolongkan sebagai positif iaitu kelas multimedia di dalam ruang ciri. Lokasi titik datanya mesti jauh dari sempadan kelas di bahagian positif. Sebaliknya, satu dokumen lain yang merangkumi banyak ciri yang tidak berkait dengan multimedia sepatutnya ditempatkan jauh dari sempadan kelas di bahagian negatif (Yu et al. 2002).



RAJAH 2. Perwakilan grafik satu SVM linear dalam kes dua dimensi
Sumber: Yu et al. 2002

SVM mempunyai satu parameter, *C* yang digunakan bagi kes-kes di mana titik-titik data tidak dapat diasingkan secara *linear*. Parameter ini merupakan penalti yang dikenakan ke atas data latihan yang jatuh ke bahagian yang salah dalam sempadan kelas. SVM akan menghitung hipersatah yang memaksimumkan jarak ke *support vectors* bagi satu nilai parameter yang diberi. Masalah-masalah yang tidak dapat diasingkan secara *linear* dapat diselesaikan dengan menggunakan kaedah kernel lanjutan yang berfungsi untuk mengubah satu ruang input bukan-linear kepada ruang ciri linear. Kaedah kernel lanjutan yang terdapat dalam SVM ialah polinomial, fungsi radial basis (RBF) dan sigmoid tanh (Lin et al. 2003).

JUSTIFIKASI PEMILIHAN KAEDAH PENGELASAN SVM

Pemilihan penggunaan kaedah pengelasan SVM dalam kajian ini adalah didorong oleh pelbagai faktor seperti berikut:

(i) Pemilihan perkataan ataupun ciri yang spesifik untuk pembelajaran adalah tidak perlu kerana SVM merupakan algoritma yang tegap dengan perlindungan kepada masalah *overfitting* dan boleh menampung dimensi yang tinggi (Sebastiani 2002).

(ii) SVM boleh menampung jumlah ciri-ciri yang banyak sehingga lebih daripada 10,000 dokumen untuk proses pembelajaran.

(iii) Kebanyakan masalah dalam pengelasan teks, khususnya kategori boleh diasingkan dan diselesaikan secara *linear*. Walau bagaimanapun, terdapat kategori tidak dapat diasingkan secara linear; disebabkan oleh dokumen bagi set data tersebut mempunyai banyak kandungan yang kurang bermakna. Bagi menangani kewujudan masalah *linear* ini, SVM boleh mencari pembahagi yang sesuai dengan menggunakan kaedah kernel lanjutan (Joachims 1998).

Selain itu, SVM tidak memerlukan tenaga manusia atau tenaga mesin dalam menentukan nilai parameter yang sesuai kerana SVM telah menyediakan set parameter piawai yang telah terbukti berkesan dalam pengelasan teks (Sebastiani 2002).

METODOLOGI

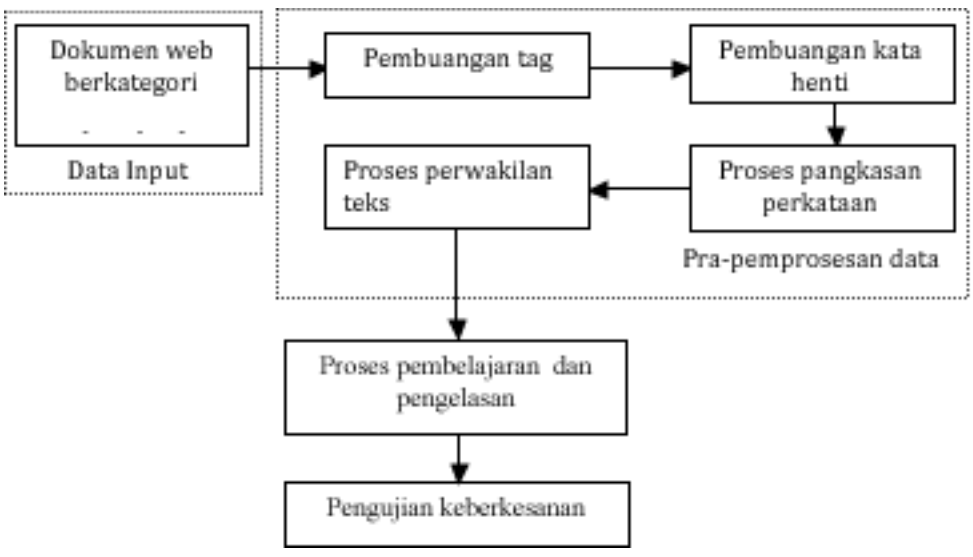
Kajian ini merupakan kajian kuantitatif dalam bidang pengelasan teks yang melibatkan gabungan teknik capaian maklumat dan kaedah pembelajaran mesin, iaitu *Support Vector Machine* (SVM). Kajian yang dilaksanakan ini adalah bertujuan untuk menguji keberkesanan kaedah pengelasan yang dipilih dengan menilai peratus ketepatan dan dapatan semula bagi proses pengelasan yang dilakukan.

Set data yang digunakan dalam kajian ini adalah laman web syarikat yang terdapat di Papan Utama dan Kedua, Bursa Malaysia. Set data yang digunakan dikategorikan mengikut jenis perniagaan yang dijalankan seperti barangan pengguna, barangan industri dan pembinaan. Jadual 2 menunjukkan jumlah laman web syarikat mengikut kategori yang terdapat di Papan Utama dan Kedua, Bursa Malaysia.

JADUAL 2. Bilangan laman web syarikat mengikut kategori di Bursa Malaysia

Bidang	Papan Utama	Papan Kedua	Jumlah
Barangan Industri	154	130	284
Dagangan / khidmat	138	48	186
Barangan Pengguna	93	50	143
Hartanah	98	3	101
Pembinaan	43	16	59
Kewangan	48	0	48
Perladangan	42	4	46
Teknologi	17	6	23
Infrastruktur	9	0	9
Reits	7	0	7
Hotel	5	0	5
Dana tertutup	2	0	2
Dana pertukaran dagangan	1	0	1
Perlombongan	1	0	1
Jumlah	**658**	**257**	**915**

Secara amnya, kajian ini dikategorikan kepada tiga bahagian, iaitu pra-pemprosesan data, pembelajaran serta pengelasan dan pengujian. Rajah 3 menunjukkan reka bentuk kajian yang dijalankan.



RAJAH 3. Reka bentuk kajian

PRA PEMPROSESAN DATA

Terdapat empat aktiviti yang terlibat di dalam perlaksanaan pra pemprosesan data seperti berikut:

(i) Pembuangan tag

Proses ini merupakan proses membuang semua tag dan skrip HTML yang ada dalam dokumen web bagi mendapatkan kandungan teks sahaja. Proses penghapusan ini perlu dilaksanakan kerana proses pengelasan dokumen web hanya menitik beratkan kandungan teks sebenar dalam dokumen web. Selepas proses penghapusan tag dan skrip dilakukan, dokumen web diwakilkan sebagai dokumen yang tidak berstruktur (Embley et al. 1998).

(ii) Pembuangan kata henti

Proses ini merupakan proses menghapuskan kata henti seperti kata sandang (*a*, *an*, *the*), sendi nama (*in*, *of*, *at*), kata penghubung (*and*, *or*, *nor*) dan lain-lain daripada dokumen web yang telah melalui proses penghapusan tag dan skrip (Guo et al. 2002). Proses ini dilakukan dengan bantuan daripada satu senarai kata henti. Perbandingan dibuat di antara dokumen web, melalui proses penghapusan tag dan skrip dengan senarai kata henti tersebut bagi tujuan penghapusan kata henti. Matlamat utama proses penghapusan kata henti ialah untuk menyingkirkan hingar daripada dokumen web (Mittermayer et al. 2001).

(iii) Proses pangkasan perkataan

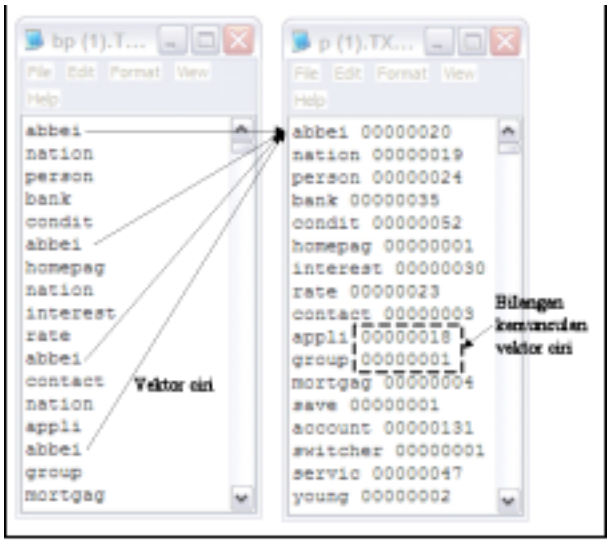
Proses pangkasan perkataan adalah proses membuang imbuhan akhiran pada setiap perkataan dalam dokumen untuk menjadikan perkataan tersebut sebagai kata akar. Contoh kata akar ialah perkataan *connect* yang telah dipangkas daripada perkataan *connected*, *connecting* dan *connections* (Baeza-Yates & Ribeiro-Neto 1999). Proses pangkasan perkataan boleh meningkatkan ketepatan proses pengelasan (Liao et al. 2001). Dalam kajian ini, proses pangkasan perkataan dilakukan dengan menggunakan algoritma Porter yang menggunakan satu senarai imbuhan akhiran untuk proses pangkasan. Algoritma ini mengaplikasikan satu siri peraturan kepada imbuhan akhiran perkataan dalam teks (Baeza-Yates & Ribeiro-Neto 1999). Dua contoh aplikasi algoritma adalah seperti berikut:

$$sses \rightarrow ss \qquad (1)$$
$$s \rightarrow \varnothing \qquad (2)$$

Berdasarkan peraturan (1), perkataan dengan aksara akhiran *sses* akan dipangkas imbuhan akhirannya iaitu *es*. Contohnya, perkataan *stresses* akan menjadi *stress* selepas melalui proses pangkasan. Berdasarkan peraturan (2) pula, perkataan majmuk yang mempunyai imbuhan akhiran *s* akan dipangkas menjadi kata akar dengan membuang aksara *s* tersebut (Baeza-Yates & Ribeiro-Neto 1999).

(iv) Proses perwakilan teks

Proses perwakilan teks merupakan proses yang terakhir dalam perlaksanaan pra-pemprosesan data. Proses ini diperlukan untuk menukar setiap perkataan tk dalam set data latihan kepada format yang difahami oleh SVM untuk dijadikan sebagai input kepada proses pembelajaran SVM. Proses perwakilan teks ini menggunakan perwakilan nilai atribut di mana setiap perkataan tk bersamaan dengan satu vektor ciri. Perwakilan teks sebagai vektor ciri boleh dilaksanakan dengan mengira bilangan kemunculan perkataan tk dalam dokumen web. Rajah 4(a) menunjukkan fail input yang memaparkan kandungan dokumen web dan Rajah 4(b) pula, adalah fail output yang memaparkan bilangan kemunculan bagi setiap vektor ciri yang terkandung dalam fail input iaitu setelah pengiraan kemunculan vektor ciri dilakukan. Contohnya, perkataan abbei muncul sebanyak 20 kali dan perkataan nation pula muncul sebanyak 19 kali sahaja. Pengiraan kemunculan vektor ciri akan dilakukan ke atas setiap dokumen web yang terlibat.



(a) Fail *input*

(b) Fail *output* (setelah pengiraan kemunculan vektor ciri dilakukan)

RAJAH 4 (a dan b). Pengiraan kemunculan vektor ciri dalam dokumen web

Disebabkan jumlah vektor ciri yang terlalu banyak akan mel ambatkan proses pembelajaran SVM (Joachims 1998); maka hanya perkataan yang muncul sekurang-kurangnya tiga kali di dalam keseluruhan set data latihan, akan diambil kira sebagai vektor ciri dan diberi nombor ciri. Selepas vektor-vektor ciri yang dipilih diberikan nombor ciri yang sepadan, setiap vektor ciri t_k yang muncul dalam dokumen d_j akan diberi nilai pemberat dengan menggunakan kaedah capaian maklumat piawai *tfidf* (Sebastiani 2002). Kaedah *tfidf* ini merupakan gabungan kaedah pengiraan frekuensi terma (*tf*) dan frekuensi dokumen songsang (*idf*). Pengiraan nilai pemberat untuk setiap vektor ciri t_k yang muncul dalam dokumen d_j adalah seperti formula berikut:

$$tfidf\,(t_k, d_j) = \#\,(t_k, d_j).\log \quad \frac{|T_r|}{\#T_r(t_k)} \qquad (3)$$

di mana:

$\#, (tdkf)$ = bilangan lemunculan vektor ciri tk dalam dokumen d_j

$\#T_r(t_k)$ = bilangan dokumen yang mempunyai vektor ciri t_k

$|T_r|$ = jumlah keseluruhan dokumen latihan

Setelah pengiraan *tfidf* dilakukan dan selepas semua vektor ciri dalam semua dokumen diberikan nilai pemberat, proses pelabelan tanda positif atau negatif untuk setiap dokumen pula dilakukan. Label positif menandakan dokumen dalam kategori c_i dan label negatif menandakan dokumen bukan dalam kategori c_i. Dokumen yang dilabelkan disusun dalam satu fail dan fail ini akan diinputkan kepada SVM untuk proses pembelajaran.

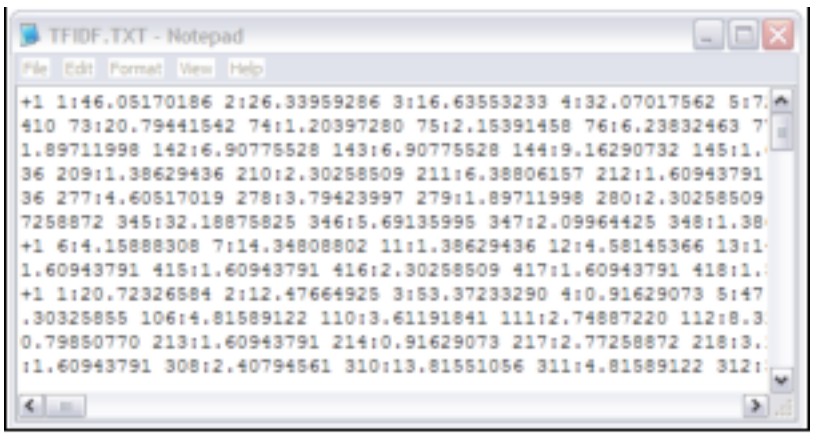
Rajah 5 merupakan format fail yang difahami oleh SVM. Merujuk kepada Rajah 5, [label] merupakan label dokumen sama ada bernilai positif atau negatif. [label] merujuk kepada kelas untuk sesuatu dokumen web. [indeks] pula merupakan nombor ciri untuk sesuatu vektor ciri yang spesifik dan bernilai integer. Manakala [nilai] merupakan nilai data yang telah diberikan pemberat untuk vektor ciri dan bernilai perpuluhan. Rajah 6 menunjukkan contoh fail yang telah ditukarkan ke format SVM.

```
[label] [indeks1]:[nilai1] [indeks2]:[nilai2]...
[label] [indeks1]:[nilai1] [indeks2]:[nilai2]...
```

RAJAH 5. Format fail SVM

Walaupun begitu, atribut pada data sebenar berkemungkinan berada di dalam julat yang sangat besar ataupun kecil. Contohnya, dalam Rajah 6, nilai minimum adalah 0.43078292 dan nilai maksimum adalah 757.95026531. Jadi, atribut tersebut perlu diskalakan semula atau dinormalkan. Tujuan penormalan data adalah untuk membolehkan data dilatih dan diramalkan dengan lebih pantas. Lazimnya, atribut data akan dinormalkan kepada skala [0,1] atau [-1,1]. Para penyelidik menggunakan skala [-1,1] dan formulanya adalah seperti berikut:

$$\text{Skala data} = \frac{x_i \; 2 \min}{\max 2 \min} \qquad (4)$$



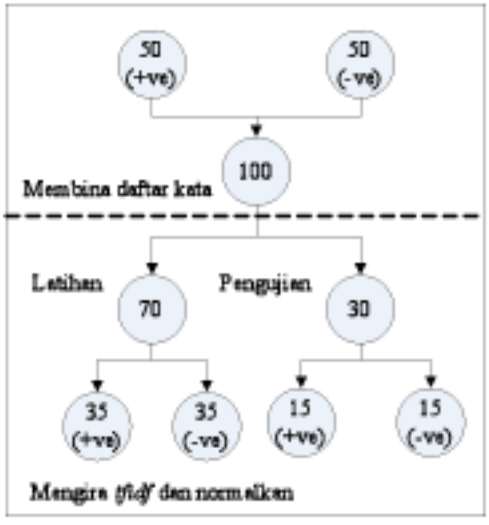
RAJAH 6. Fail yang telah ditukarkan ke format SVM

PEMBELAJARAN DAN PENGELASAN MENGGUNAKAN SVM

Dokumen web yang telah ditukarkan kepada format SVM dan yang telah melalui proses penormalan data akan melalui proses pembelajaran dan pengelasan menggunakan SVM. Proses pembelajaran dan pengelasan melibatkan dua jenis data iaitu data latihan dan pengujian. Set data pengujian terdiri daripada satu nilai sasaran (label kelas) dan atributnya iaitu ciri dokumen tersebut. Perlaksanaan proses pembelajaran set data latihan dan pengelasan set data pengujian dilakukan dengan menggunakan pakej LibSVM yang dihasilkan oleh Chang dan Lin (2000). Set data yang telah dinormalkan diinputkan kepada pakej SVM untuk proses pembelajaran. Rajah 7 menunjukkan proses pembelajaran dan pengelasan menggunakan SVM.

Sebelum proses pembelajaran dan pengelasan boleh dilaksanakan, dokumen web akan dibahagikan kepada dua set data iaitu 70% sebagai set data latihan untuk proses pembelajaran dan 30% lagi sebagai set data pengujian untuk proses pengujian pengelasan. Rajah 8 menunjukkan contoh pembahagian set data di mana jumlah data web yang digunakan adalah 100 dokumen web. Dokumen web ini terdiri daripada dua iaitu 50 dokumen web positif (dokumen web yang berada dalam kategori yang ditetapkan) dan 50 dokumen web adalah negatif (dokumen web yang bukan daripada kategori yang ditetapkan). Kegunaan dokumen ini adalah untuk membina fail daftar kata. Setelah itu, dokumen ini akan dipecahkan kepada 70 dokumen untuk latihan dan 30 dokumen pula untuk pengujian. Dokumen untuk latihan meliputi 35 dokumen positif dan negatif. Manakala dokumen untuk pengujian pula, merangkumi 15 dokumen positif dan negatif.

RAJAH 7. Proses pembelajaran dan pengelasan menggunakan SVM



RAJAH 8. Pembahagian set data untuk latihan dan pengujian

Dalam kajian ini hanya enam bidang sahaja terlibat iaitu barangan industri, dagangan ataupun khidmat, barangan pengguna, hartanah, pembinaan dan kewangan. Bidang-bidang lain seperti infrastruktur, perlombongan dan teknologi tidak digunakan untuk proses pengelasan web kerana bidang-bidang tersebut mempunyai dokumen web yang terlalu sedikit iaitu antara 8 hingga 17 dokumen sahaja. Maklumat berkaitan set data yang digunakan untuk pengujian ketepatan pengelasan web boleh dirujuk dalam Jadual 3. Setelah pembahagian data dilaksanakan, LibSVM digunakan untuk membolehkan data dikelaskan.

Di samping itu, dokumen-dokumen yang digunakan untuk pengujian, mempunyai format yang sama dengan dokumen-dokumen latihan. Dokumen-dokumen pengujian yang telah melalui semua aktiviti dalam pra-pemprosesan data ini, disusun dalam satu fail pengujian mengikut label dan diinputkan ke dalam LibSVM untuk proses pengelasan. Selepas proses pengelasan selesai, LibSVM akan menjana satu fail *predictions* yang menyimpan nilai jangkaan keputusan untuk dokumen-dokumen pengujian yang telah dikelaskan. Label positif atau negatif pada nilai-nilai ini menentukan kategori satu-satu dokumen pengujian. Sekiranya nilai adalah positif, maka dokumen pengujian tersebut dijangka tergolong dalam kategori c_i. Sebaliknya, sekiranya nilai adalah negatif, maka dokumen pengujian tersebut dijangka bukan tergolong dalam kategori c_i. Set data pengujian dikelaskan untuk menguji keberkesanan dan ketepatan proses pembelajaran.

JADUAL 3. Pembahagian set data untuk latihan dan pengujian

Bidang	Data	Latihan	Pengujian
		70%	30%
Barangan Industri	156	109	47
Dagangan atau Khidmat	122	85	37
Barangan Pengguna	75	53	22
Hartanah	57	40	17
Pembinaan	35	25	11
Kewangan	35	25	11
Jumlah	**480**	**336**	**144**

PENGUJIAN KETEPATAN PENGELASAN

Bagi tujuan menguji ketepatan bagi hasil proses pengelasan menggunakan SVM, para penyelidik telah membahagikan ciri dokumen web kepada enam bahagian iaitu teks, meta tag dan tajuk (A), tajuk dan teks (B), tajuk (C), meta tag dan tajuk (D), meta tag (E) dan teks (E). Tujuan pembahagian ciri dokumen web ini dilakukan adalah bagi mengenal pasti ciri yang terbaik untuk mengelaskan dokumen web. Penerangan lanjut berkaitan dengan ciri dokumen web ini adalah seperti berikut:

(i) Teks, meta tag dan tajuk (A)
Merangkumi maklumat yang terdiri daripada teks, meta tag dan tajuk.

(ii) Tajuk dan teks (B)
Meliputi maklumat yang terdapat pada tag <body> hingga </body> dan tag tajuk sahaja. Maklumat pada meta tag lain akan diabaikan.

(iii) Tajuk (C)
Maklumat yang diambil daripada tajuk dokumen web. Elemen tajuk ini adalah berdasarkan nilai pada <title> hingga </title>.

(iv) Meta tag dan tajuk (D)
Merangkumi maklumat meta tag dan tajuk sahaja.

(v) Meta tag (E)
Meta tag ini hanya melibatkan tag kata kunci dan penerangan. Elemen meta tag ini berdasarkan nilai yang diambil dari <meta name="description" content=""> dan <meta name="keywords" content="">.

(vi) Teks sahaja (F)
Meliputi maklumat teks yang terdapat pada tag <body> hingga </body> sahaja. Maklumat pada meta tag lain akan diabaikan.

Selain itu, bagi melaksanakan pengujian ketepatan pengelasan web ini, teknik pengesahan silang 5 dan 10 lipatan serta empat jenis kernel iaitu kernel fungsi radial basis (RBF), linear, polinomial dan sigmoid akan digunakan. Pengujian ketepatan proses pengelasan ini juga akan menggunakan kaedah pengukuran pencapaian statistik iaitu nilai ketepatan. Nilai ketepatan didefinisikan sebagai kebarangkalian satu dokumen yang dijangkakan tergolong dalam kategori c_i adalah benar-benar tergolong dalam kategori c_i (Joachims 1998). Berikut adalah formula bagi mengira nilai ketepatan (Yang 1998):

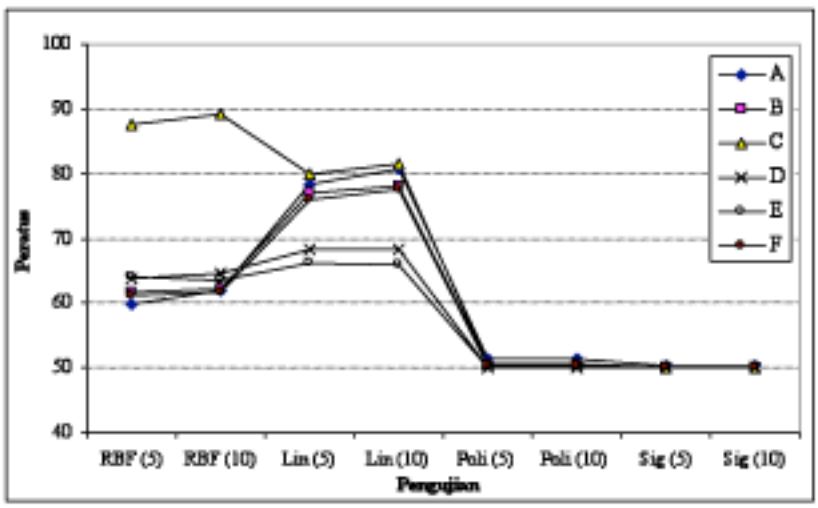
$$\text{Ketepatan} = \frac{\text{Jumlah dokumen relevan yang dicapai}}{\text{Jumlah dokumen yang dicapai}} \qquad (5)$$

PERBINCANGAN HASIL

Hasil pengujian ketepatan pengelasan web menunjukkan pengelasan web berdasarkan ciri tajuk (C) mencatatkan purata peratusan ketepatan tertinggi berbanding pengelasan web menggunakan ciri yang lain. Kajian juga turut mendapati purata peratusan ketepatan pengelasan web menggunakan kernel *linear* adalah lebih baik berbanding yang lain di mana purata peratusan ketepatan pengelasan webnya melebihi 65%. Pengelasan web menggunakan kernel RBF hanya mencatatkan purata peratusan ketepatan tertinggi apabila pengelasan dibuat berdasarkan ciri tajuk dan selebihnya pengelasan web menggunakan kernel linear akan memberikan peratusan ketepatan yang lebih baik. Penggunaan kernel polinomial dan sigmoid mencatatkan peratusan ketepatan yang tidak memberansangkan iaitu hanya dalam lingkungan 50%.

Kajian mengikut ciri dokumen web pula, menunjukkan pengujian C iaitu ciri tajuk mencatatkan peratusan pengelasan tertinggi dengan 81.50% dan diikuti oleh pengujian A iaitu ciri teks, meta tag dan tajuk yang mencatatkan peratusan sebanyak 80.61%. Selain itu, terdapat dua pengujian yang menghasilkan peratus ketepatan pengelasan di antara 70% sehingga 80% iaitu pengujian B sebanyak 78.19% dan pengujian F pula sebanyak 77.60%. Kajian mendapati pengujian ke atas penggunaan ciri meta tag dan tajuk (D) dan pengujian ciri meta tag (E) mencatatkan peratusan di bawah 70% iaitu 68.27% untuk pengujian D dan 66.21% untuk pengujian E. Rajah 9 dan Jadual 4 menunjukkan purata peratusan pengelasan web mengikut ciri pengujian.

Petunjuk:
RBF - Fungsi radial basis,
Poli - Polinomial,
Sig - Sigmoid,
Lin - Linear,
A - Teks, meta tag dan tajuk,
B - Tajuk dan teks,
C - Tajuk,
D - Meta tag dan tajuk,
E - Meta tag,
F - Teks



RAJAH 9. Purata peratusan pengelasan web mengikut kernel dan pengesahan silang

JADUAL 4. Purata peratusan pengelasan web mengikut ciri pengujian

Pengujian	RBF		Linear		Polinomial		Sigmoid	
	5 lip	10 lip	5 lip	10 lip	5 lip	10 lip	5 lip	10 lip
A - teks, meta tag dan tajuk	59.77	61.87	78.44	80.61	51.49	51.43	50.34	50.24
B - tajuk dan teks	61.70	62.10	76.91	78.19	50.48	50.49	50.00	50.00
C – tajuk	87.59	89.05	79.83	81.50	50.54	50.54	50.00	50.00
D - meta tag dan tajuk	63.68	64.57	68.18	68.27	50.00	50.00	50.00	50.00
E - meta tag	63.97	63.61	66.21	66.01	50.00	50.00	50.00	50.00
F – teks	61.13	61.74	75.85	77.60	50.33	50.41	50.00	50.00

Petunjuk: lip - lipatan

KESIMPULAN

Secara keseluruhannya, didapati bahawa penggunaan ciri tajuk (C) mampu memberikan peratusan ketepatan pengelasan dokumen web yang tinggi berbanding ciri-ciri dokumen web yang lain. Penggunaan kernel yang terbaik bagi pengelasan web pula, adalah kernel *linear* dengan peratusan ketepatan adalah 89.80%. Hasil kajian ini menyokong kajian yang dijalankan oleh Hsu dan Lin (2002) ke atas set data yang berbeza seperti set data *iris*, *wine* dan *segments* yang menunjukkan penggunaan kernel *linear* menghasilkan pengelasan yang terbaik diikuti dengan kernel RBF. Manakala kajian yang dilaksanakan oleh Wang et al. (2005) ke atas set data institusi kewangan Taiwan dan bank komersial Amerika Syarikat turut menghasilkan dapatan hasil pengelasan yang sama.

RUJUKAN

Amoretti, M.S.M. 2006. Categorization Process and Data Mining. Dlm. Wang, J. *Encyclopedia of Data Warehousing and Mining*, hlm. 129-133. London: Idea Group Inc.

Baeza-Yates, R. & Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Essex: Addison Wesley.

Chang, C-C. & Lin, C.J. 2000. LIBSVM: an integrated software for support vector classification and regression. (atas talian) http://www.csie.ntu.edu.tw/~cjlin/ libsvm/ (4 Jun 2005).

Cortes, C. & Vapnik, V. 1995. Support-vector networks. *Machine Learning* 20: 273-297.

Drucker, H., Vapnik, V. & Wu, D. 1999. Automatic Text Categorization and Its Applications to Text Retrieval. *IEEE Trans. Neural Network* 10(5): 1048-1054.

Dumais, S. T., Platt, J., Heckerman, D., & Sahami, M. 1998. Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge*, hlm. 148-155.

Embley, D. W., Campbell, D. M., Jiang, Y. S., Ng, Y. K. dan Smith, R. D. (1998). *A Conceptual-Modeling Approach to Extracting Data for the Web*. Proceedings of the 17th International Conference on Conceptual Modeling (ER'98), hlm 78-91

Etzioni, O. 1996. The World Wide Web: quagmire or gold mine?. *Communications of the ACM* 39(11):65-68.

Gulli, A. & Signorini, A. 2005. The indexable web is more than 11.5 billion pages. *International World Wide Web Conference- Special interest tracks and posters of the 14th International Conference on World Wide Web*, hlm. 902-903.

Guo G, Wang H, Bell D, Bi Y & Greer Y. 2004. Using kNN Model for Automatic Text Categorization. *Journal of Soft Computing* 2888:986-996.

Hammerich, I, dan Harrison, C. 2002, "*Developing Online Content : The Principles of Writing and Editing for the Web*", John Wiley & Sons Inc, USA

Henzinger, M.R., Motwani, R., Silverstein, C.2003. Challenges in Web Search Engines. Proc. of the 18th International Joint Conference on Artificial Intelligence, hlm. 1573-1579.

Hizral Tazzif Hisham. 2005. Pelaksanaan E-Dagang Hadapi Cabaran Besar. *Berita Harian***.** 21 November: 15.

Huang, L. 2000. A survey on web information retrieval technologies. *Laporan Teknik ECSL*. (atas talian) http://www.ecsl.cs.sunysb.edu/tr/rpe8.ps.Z. (5 Mei 2004)

Joachims, T. 1998. Text Categorization with Suport Vector Machines: Learning with Many Relevant Features, *Proceedings of the 10th European Conference on Machine Learning*, hlm.137-142.

Klinkenberg, R. & Joachims, T. 2000. Detecting Concept Drift With Support Vector Machines. Proceedings of ICML-00, 17th International Conference on Machine Learning, hlm. 487-494.

Kosala, R. & Blockeel, H. 2000. Web Mining Research: A Survey. *ACM SIGKDD* 2(1): 1-15.

Lawrence, S. & Giles, C. L. 1999. Accessibility of information on the web. *Nature* 400: 107-109.

Liao, C., Alpha, S. & Dixon, P. 2001. Feature Preparation in Text Categorization. The Australian Data Mining Workshop, Canberra, Australia.

Lin, C. J., Hsu, C. W. & Chang, C. C. 2003. A Practical Guide to Support Vector Classification. National Taiwan University, Taipei 106, Taiwan. (atas talian) http://www. csie.ntu.edu.tw/~cjlin/papers.html (21 Oktober 2006).

Maron, M. 1961. Automatic indexing: An experimental inquiry. *Journal of the Association for Computing Machinery* 8(3): 404-417.

Mittermayer, M. A., Brucher, H. & Knolmayer, G. 2001. Document Classification Methods for Organizing Explicit Knowledge. *Proceedings of the Third European Conference on Organizational Knowledge, Learning and Capabilities*, Athens.

Mohd Shahizan Othman , Lizawati Mi Yusuf, Juhana Salim dan Zarina Shukur. 2005. Kajian Terhadap Penggunaan Tag, Meta Tag Dan Perkataan Bagi Sumber Maklumat Web . *Simposium Kebangsaan Sains Matematik ke 13*. Hotel Holiday Villa, Alor Star. 31 Mei - 2 Jun. Jil. 2:975-984

Pierre, J. M. 2001. On the automated classification of web sites. *Electronic Transactions on Artificial Intelligence*. (dalam talian) http://www.ida.liu.se/ext/etai/ra/seweb/ 002/ (01 Jun 2006).

Rachagan, S. 2005. Rakyat tidak boleh harapkan usaha kerajaan tapis internet. *Berita Harian*, 11 Ogos:10.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. A*CM Computing Surveys,* 34(1): 1-47.

Sebastiani, F. 2005. Text Categorization. Dlm. Alessandro Zanasi (pnyt.). *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*, hlm. 109-129. Southampton: WIT Press.

Wang, Y., Wang, S. dan Lai, K.K. 2005. A New Fuzzy Support Vector Machine to Evaluate Credir Risk. IEEE Transactions on Fuzzy Systems 13(6): 820-831

Yang, Y. 1998. *An Evaluation of Statistical Approaches to Text Categorization*. The Netherlands: Kluwer Academic Publishers. 69-90

Yang, Y. & Liu, X. 1999. A Re-examination of Text Categorization Methods. *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, hlm. 42-49.

Yu, H., Han, J. & Chang, K. C. C. 2002. PEBL: Positive Example Based Learning for Web Page Classification Using SVM. *Proc. ACM SIGKDD International Conference Knowledge Discovery in Databases*, hlm 239-248.

Mohd Shahizan Othman & Lizawati Mi Yusuf
Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia
81310 Skudai, Johor Bahru
Johor
shahizan@utm.my
lizawati@utm.my

Juhana Salim & Zarina Shukur
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia
43600 UKM Bangi
Selangor
js@ftsm.ukm.my
zs@ftsm.ukm.my